Lecture 8: Inference
36-401, Fall 2018, Section B

Having gone over the Gaussian-noise simple linear regression model, over ways of estimating its parameters and some of the properties of the model, and over how to check the model's assumptions, we are now ready to begin doing some serious statistical inference within the model. In previous lectures, we came up with **point estimators** of the parameters and the conditional mean (prediction) function, but we weren't able to say much about the margin of uncertainty around these estimates. In this lecture we will focus on supplementing point estimates with *reliable* measures of uncertainty. This will naturally lead us to testing hypotheses about the true parameters — again, we will want hypothesis tests which are unlikely to get the answer wrong, whatever the truth might be.

To accomplish all this, we first need to understand the sampling distribution of our point estimators. We can find them, mathematically, but they involve the unknown true parameters in inconvenient ways. We will therefore work to find combinations of our estimators and the true parameters with fixed, parameter-free distributions; we'll get our confidence sets and our hypothesis tests from them.

Throughout this lecture, I am assuming, unless otherwise noted, that all of the assumptions of the Gaussian-noise simple linear regression model hold.

1 Sampling Distribution of $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\sigma}^2$

The Gaussian-noise simple linear regression model has three parameters: the intercept β_0, the slope β_1, and the noise variance σ^2. We've seen, previously, how to estimate all of these by maximum likelihood; the MLE for the βs is the same as their least-squares estimates. These are

$$\widehat{\beta}_1 = \frac{c_{XY}}{s_X^2} = \sum_{i=1}^{n} \frac{X_i - \overline{x}}{n s_X^2} Y_i \tag{1}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{2}$$

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i))^2 \tag{3}$$

We have also seen how to re-write the first two of these as a deterministic part plus a weighted sum of the noise terms ϵ:

$$\widehat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \frac{X_i - \overline{x}}{n s_X^2} \epsilon_i \tag{4}$$

$$\widehat{\beta}_0 = \beta_0 + \frac{1}{n} \sum_{i=1}^{n} \left(1 - \overline{x} \frac{X_i - \overline{x}}{s_X^2}\right) \epsilon_i \tag{5}$$

Finally, we have our modeling assumption that the ϵ_i are independent Gaussians, $\epsilon_i \sim N(0, \sigma^2)$.

1.1 Reminders of Basic Properties of Gaussian Distributions

Suppose $U \sim N(\mu, \sigma^2)$. By the basic algebra of expectations and variances, $\mathbb{E}[a + bU] = a + b\mu$, while $\mathrm{Var}[a + bU] = b^2\sigma^2$. This would be true of any random variable; a special property of Gaussians.

Suppose $U_1, U_2, \ldots U_n$ are *independent* Gaussians, with means μ_i and variances σ_i^2. Then

$$\sum_{i=1}^{n} U_i \sim N\left(\sum_i \mu_i, \sum_i \sigma_i^2\right)$$

That the expected values add up for a sum is true of all random variables; that the variances add up is true for all uncorrelated random variables. That the sum follows the same type of distribution as the summands is a special property of Gaussians.

1.2 Sampling Distribution of $\widehat{\beta}_1$

Since we're assuming Gaussian noise, the ϵ_i are independent Gaussians, $\epsilon_i \sim N(0, \sigma^2)$. Hence (using the first basic property of Gaussians)

$$\frac{X_i - \overline{x}}{n s_X^2}\epsilon_i \sim N\left(0, \left(\frac{X_i - \overline{x}}{n s_X^2}\right)^2 \sigma^2\right)$$

Thus, using the second basic property of Gaussians,

$$\begin{aligned}
\sum_{i=1}^{n} \frac{X_i - \overline{x}}{n s_X^2}\epsilon_i &\sim N\left(0, \sigma^2 \sum_{i=1}^{n}\left(\frac{X_i - \overline{x}}{n s_X^2}\right)^2\right) && (6)\\
&= N\left(0, \frac{\sigma^2}{n s_X^2}\right) && (7)
\end{aligned}$$

Using the first property of Gaussians again,

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{n s_X^2}\right) \quad (8)$$

This is the distribution of estimates we'd see if we repeated the experiment (survey, observation, etc.) many times, and collected the results. Every particular run of the experiment would give a slightly different $\widehat{\beta}_1$, but they'd average out to β_1, the average squared difference from β_1 would be $\sigma^2/n s_X^2$, and a histogram of them would follow the Gaussian probability density function (Figure 2).

It is a bit hard to use Eq. 8, because it involves two of the unknown parameters. We can manipulate it a bit to remove one of the parameters from the probability distribution,

$$\widehat{\beta}_1 - \beta_1 \sim N(0, \frac{\sigma^2}{n s_X^2})$$

but that still has σ^2 on the right hand side, so we can't actually calculate anything. We could write

$$\frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{n s_X^2}} \sim N(0,1)$$

but now we've got two unknown parameters on the left-hand side, which is also awkward.

```
# Simulate a Gaussian-noise simple linear regression model
# Inputs: x sequence; intercept; slope; noise variance; switch for whether to
  # return the simulated values, or run a regression and return the coefficients
# Output: data frame or coefficient vector
sim.gnslrm <- function(x, intercept, slope, sigma.sq, coefficients=TRUE) {
  n <- length(x)
  y <- intercept + slope*x + rnorm(n,mean=0,sd=sqrt(sigma.sq))
  if (coefficients) {
    return(coefficients(lm(y~x)))
  } else {
    return(data.frame(x=x, y=y))
  }
}

# Fix an arbitrary vector of x's
x <- seq(from=-5, to=5, length.out=42)
```

FIGURE 1: *Code setting up a simulation of a Gaussian-noise simple linear regression model, along a fixed vector of X_i values.*

1.3 Sampling Distribution of $\widehat{\beta}_0$

Starting from Eq. 5 rather than Eq. 4, an argument exactly parallel to the one we just went through gives

$$\widehat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_X^2}\right))$$

It follows, again by parallel reasoning, that

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_X^2}\right)}} \sim N(0,1)$$

The right-hand side of this equation is admirably simple and easy for us to calculate, but the left-hand side unfortunately involves two unknown parameters, and that complicates any attempt to use it.

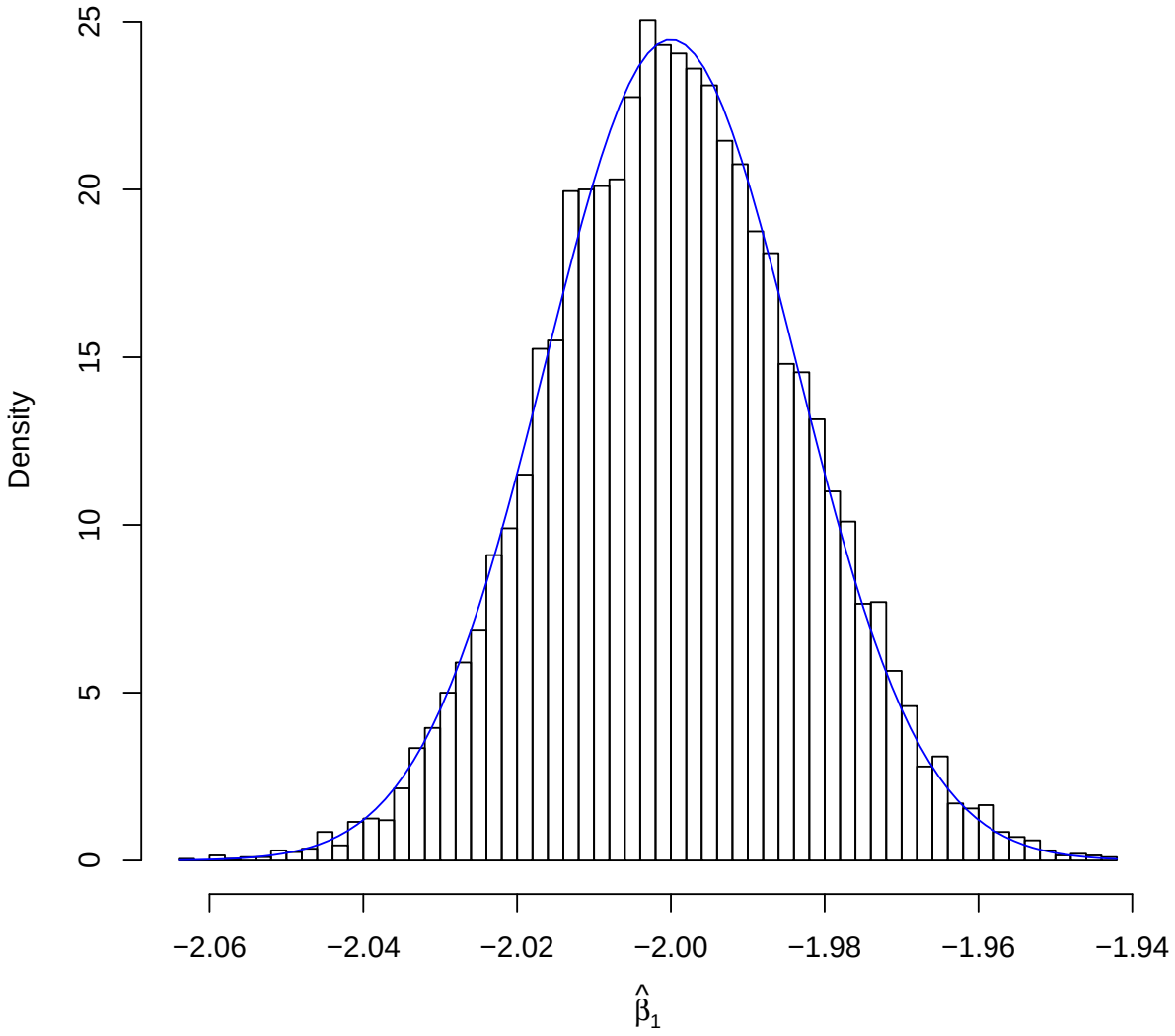
1.4 Sampling Distribution of $\widehat{\sigma}^2$

It is mildly challenging, but certainly not too hard, to show that

$$\mathbb{E}\left[\widehat{\sigma}^2\right] = \frac{n-2}{n}\sigma^2$$

We can be much more specific. When $\epsilon_i \sim N(0, \sigma^2)$, it can be shown that

$$\frac{n\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$




```
# Run the simulation 10,000 times and collect all the coefficients
  # What intercept, slope and noise variance does this impose?
many.coefs <- replicate(1e4, sim.gnslrm(x=x, 5, -2, 0.1, coefficients=TRUE))
# Histogram of the slope estimates
hist(many.coefs[2,], breaks=50, freq=FALSE, xlab=expression(hat(beta)[1]),
     main="")
# Theoretical Gaussian sampling distribution
theoretical.se <- sqrt(0.1/(length(x)*var(x)))
curve(dnorm(x,mean=-2,sd=theoretical.se), add=TRUE,
      col="blue")
```

FIGURE 2: *Simulating 10,000 runs of a Gaussian-noise simple linear regression model, calculating* $\widehat{\beta}_1$ *each time, and comparing the histogram of estimates to the theoretical Gaussian distribution (Eq. 8, in blue).*

1.5 Standard Errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$

The **standard error** of an estimator is its standard deviation. We've just seen that the true standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are, respectively,

$$\text{se}\left[\widehat{\beta}_1\right] = \frac{\sigma}{s_x\sqrt{n}} \tag{9}$$

$$\text{se}\left[\widehat{\beta}_0\right] = \frac{\sigma}{\sqrt{n}s_X}\sqrt{s_X^2 + \overline{x}^2} \tag{10}$$

Unfortunately, these standard errors involve the unknown parameter σ^2 (or its square root σ, equally unknown to us).

We can, however, *estimate* the standard errors. The maximum-likelihood estimates just substitute $\widehat{\sigma}$ for σ:

$$\widehat{\text{se}}\left[\widehat{\beta}_1\right] = \frac{\widehat{\sigma}}{s_x\sqrt{n}} \tag{11}$$

$$\widehat{\text{se}}\left[\widehat{\beta}_0\right] = \frac{\widehat{\sigma}}{s_X\sqrt{n}}\sqrt{s_X^2 + \overline{x}^2} \tag{12}$$

For later theoretical purposes, however, things will work out slightly nicer if we use the de-biased version, $\frac{n}{n-2}\widehat{\sigma}^2$:

$$\widehat{\text{se}}\left[\widehat{\beta}_1\right] = \frac{\widehat{\sigma}}{s_x\sqrt{n-2}} \tag{13}$$

$$\widehat{\text{se}}\left[\widehat{\beta}_0\right] = \frac{\widehat{\sigma}}{s_x\sqrt{n-2}}\sqrt{s_X^2 + \overline{x}^2} \tag{14}$$

These standard errors — approximate or estimated though they be — are one important way of quantifying how much uncertainty there is around our point estimates. However, we can't use them, *alone* to say anything terribly precise about, say, the probability that β_1 is in the interval $[\widehat{\beta}_1 - \widehat{\text{se}}\left[\widehat{\beta}_1\right], \widehat{\beta}_1 - \widehat{\text{se}}\left[\widehat{\beta}_1\right]]$, which is the sort of thing we'd want to be able to give guarantees about the reliability of our estimates.

2 Sampling distribution of $(\widehat{\beta} - \beta)/\widehat{\text{se}}\left[\widehat{\beta}\right]$

It should take only a little work with the properties of the Gaussian distribution to convince yourself that

$$\frac{\widehat{\beta}_1 - \beta_1}{\text{se}\left[\widehat{\beta}_1\right]} \sim N(0,1)$$

the standard Gaussian distribution. If the Oracle told us σ^2, we'd know $\mathrm{se}\left[\widehat{\beta}_1\right]$, and so we could assert that (for example)

$$\begin{aligned}
& \mathbb{P}\left(\beta_1 - 1.96\mathrm{se}\left[\widehat{\beta}_1\right] \leq \widehat{\beta}_1 \leq \beta_1 + 1.96\mathrm{se}\left[\widehat{\beta}_1\right]\right) && (15)\\
= \;& \mathbb{P}\left(-1.96\mathrm{se}\left[\widehat{\beta}_1\right] \leq \widehat{\beta}_1 - \beta_1 \leq 1.96\mathrm{se}\left[\widehat{\beta}_1\right]\right) && (16)\\
= \;& \mathbb{P}\left(-1.96 \leq \frac{\widehat{\beta}_1 - \beta_1}{\mathrm{se}\left[\widehat{\beta}_1\right]} \leq 1.96\right) && (17)\\
= \;& \Phi(1.96) - \Phi(-1.96) = 0.95 && (18)
\end{aligned}$$

where Φ is the cumulative distribution function of the $N(0,1)$ distribution.

Since the oracles have fallen silent, we can't use this approach. What we *can* do is use the following fact:

Proposition 1 *If $Z \sim N(0,1)$, $S^2 \sim \chi^2_d$, and Z and S^2 are independent, then*

$$\frac{Z}{\sqrt{S^2/d}} \sim t_d$$

(I call this a proposition, but it's almost a definition of what we mean by a t distribution with d degrees of freedom. Of course, if we take this as the definition, the proposition that this distribution has a probability density $\propto (1 + x^2/d)^{-(d+1)/2}$ would become yet another proposition to be demonstrated.)

Let's try to manipulate $(\widehat{\beta}_1 - \beta_1)/\widehat{\mathrm{se}}\left[\widehat{\beta}_1\right]$ into this form.

$$\begin{aligned}
\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathrm{se}}\left[\widehat{\beta}_1\right]} &= \frac{\widehat{\beta}_1 - \beta_1}{\sigma}\frac{\sigma}{\widehat{\mathrm{se}}\left[\widehat{\beta}_1\right]}\\
&= \frac{\frac{\widehat{\beta}_1 - \beta_1}{\sigma}}{\frac{\widehat{\mathrm{se}}\left[\widehat{\beta}_1\right]}{\sigma}} = \frac{N(0, 1/ns_X^2)}{\frac{\widehat{\sigma}}{s_x\sigma\sqrt{n-2}}} = \frac{s_X N(0, 1/ns_X^2)}{\frac{\widehat{\sigma}}{\sigma\sqrt{n-2}}}\\
&= \frac{N(0,1/n)}{\frac{\widehat{\sigma}}{\sigma\sqrt{n-2}}} = \frac{\sqrt{n}N(0,1/n)}{\frac{\sqrt{n}\widehat{\sigma}}{\sigma\sqrt{n-2}}} = \frac{N(0,1)}{\sqrt{\frac{n\widehat{\sigma}^2}{\sigma^2}\frac{1}{n-2}}}\\
&= \frac{N(0,1)}{\sqrt{\chi^2_{n-2}/(n-2)}} = t_{n-2}
\end{aligned}$$

where in the last step I've used the proposition I stated (without proof) above.

To sum up:

Proposition 2 *Using the $\widehat{\mathrm{se}}\left[\widehat{\beta}_1\right]$ of Eq. 13,*

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathrm{se}}\left[\widehat{\beta}_1\right]} \sim t_{n-2} \tag{19}$$

Notice that we can compute $\widehat{\text{se}}\left[\widehat{\beta}_1\right]$ without knowing any of the true parameters — it's a pure statistic, just a function of the data. This is a key to actually using the proposition for anything useful.

By exactly parallel reasoning, we may also demonstrate that

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\text{se}}\left[\widehat{\beta}_0\right]} \sim t_{n-2}$$

3 Confidence Intervals and Tests

Define $k \equiv k(n, \alpha)$ such that

$$\int_{-k(n,\alpha)}^{k(n,\alpha)} f(u)du = 1 - \alpha$$

where f is the density of a t_{n-2} distribution. Let $s = \widehat{\text{se}}(\widehat{\beta}_1)$. A $1 - \alpha$ confidence interval for β_1 is

$$C = \left[\widehat{\beta}_1 - ks,\ \widehat{\beta}_1 + ks\right].$$

To verify this, note that

$$\begin{aligned} P(\beta_1 \in C) &= P(\widehat{\beta}_1 - ks \leq \beta_1 \leq \widehat{\beta}_1 + ks) = P\left(-k \leq \frac{\widehat{\beta}_1 - \beta_1}{s} < k\right) \\ &= P(-k < T < k) = 1 - \alpha \end{aligned}$$

where T denotes a random var iable with a t_{n-2} distribution. So the interval traps β_1 with probability $1 - \alpha$.

Suppose we want to test

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0.$$

We can just reject H_0 is 0 is not in C. Equivalently, reject H_0 if

$$\frac{|\widehat{\beta}_1|}{s} > k(n, \alpha).$$

This is called the **Wald test**.

Width of the confidence interval Notice that the width of the confidence interval is $2k(n, \alpha)\widehat{\text{se}}\left[\widehat{\beta}_1\right]$. This tells us what controls the width of the confidence interval:

1. As α shrinks, the interval widens. (High confidence comes at the price of big margins of error.)
2. As n grows, the interval shrinks. (Large samples mean precise estimates.)
3. As σ^2 increases, the interval widens. (The more noise there is around the regression line, the less precisely we can measure the line.)
4. As s_X^2 grows, the interval shrinks. (Widely-spread measurements give us a precise estimate of the slope.)

What about β_0? By exactly parallel reasoning, a $1-\alpha$ confidence interval for β_0 is $[\widehat{\beta}_0 - k(n,\alpha)\widehat{\text{se}}\left[\widehat{\beta}_0\right], \widehat{\beta}_0 + k(n,\alpha)\widehat{\text{se}}\left[\widehat{\beta}_0\right]]$.

What about σ^2? See Exercise 1.

What α should we use? It's become conventional to set $\alpha = 0.05$. To be honest, this owes more to the fact that the resulting k tends to 1.96 as $n \to \infty$, and $1.96 \approx 2$, and most psychologists and economists could multiply by 2, even in 1950, than to any genuine principle of statistics or scientific method. A 5% error rate corresponds to messing up about one working day in every month, which you might well find high. On the other hand, there is nothing which stops you from increasing α. It's often illuminating to plot a series of confidence sets, at different values of α.

What about power? The **coverage** of a confidence set is the probability that it includes the true parameter value. This is not, however, the only virtue we want in a confidence set; if it was, we could just say "Every possible parameter is in the set", and have 100% coverage no matter what. We would also like the *wrong* values of the parameter to have a high probability of *not* being in the set. Just as the coverage is controlled by the size / false-alarm probability / type-I error rate α of the hypothesis test, the probability of excluding the wrong parameters is controlled by the power / miss probability / type-II error rate. Test with higher power exclude (correctly) more parameter values, and give smaller confidence sets.

3.1 Confidence Sets and Hypothesis Tests

There is a general relationship between confidence sets and hypothesis tests.

1. Inverting any hypothesis test gives us a confidence set.
2. If we have a way of constructing a $1-\alpha$ confidence set, we can use it to test the hypothesis that $\beta = \beta^*$: reject when β^* is outside the confidence set, retain the null when β^* is inside the set.

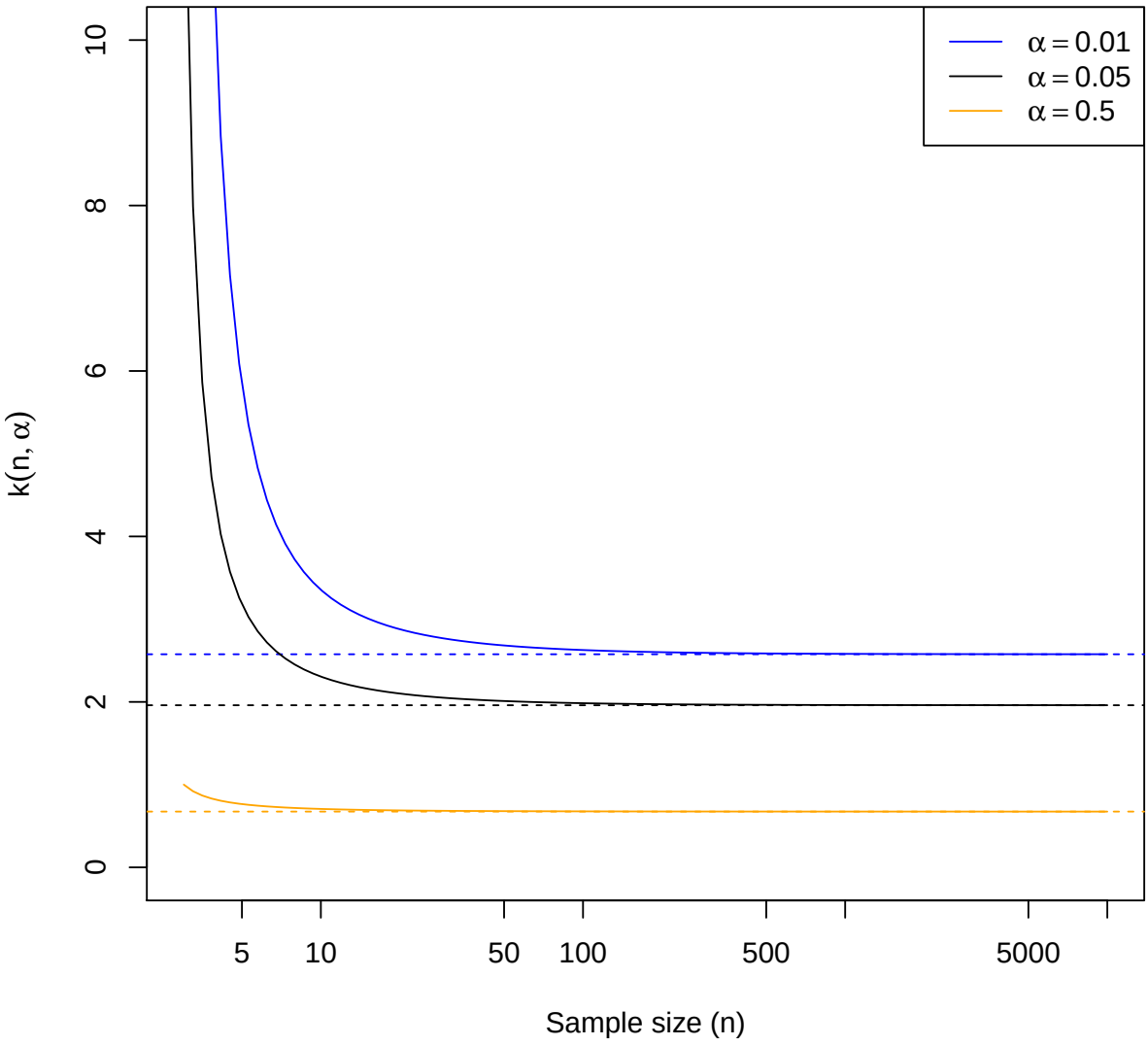
I will leave it as a pair of exercises (2 and 3) to that inverting a test of size α gives a $1-\alpha$ confidence set, and that inverting a $1-\alpha$ confidence set gives a test of size α.

3.2 Large-n Asymptotics

As $n \to \infty$, $\widehat{\sigma}^2 \to \sigma^2$. It follows (by continuity) that $\widehat{\text{se}}\left[\widehat{\beta}\right] \to \text{se}\left[\widehat{\beta}\right]$. Hence,

$$\frac{\widehat{\beta} - \beta}{\widehat{\text{se}}\left[\widehat{\beta}\right]} \to N(0,1)$$

which considerably simplifies the sampling intervals and confidence sets; as n grows, we can forget about the t distribution and just use the standard Gaussian distribution. Figure 3 plots the convergence of $k(n,\alpha)$ towards the $k(\infty,\alpha)$ we'd get from the Gaussian approximation. As you can see from the figure, by the time $n = 100$ —a quite small data set by modern standards — the difference between the t distribution and the standard-Gaussian is pretty trivial.




```
curve(qt(0.995,df=x-2),from=3,to=1e4,log="x", ylim=c(0,10),
      xlab="Sample size (n)", ylab=expression(k(n,alpha)),col="blue")
abline(h=qnorm(0.995),lty="dashed",col="blue")
curve(qt(0.975,df=x-2), add=TRUE)
abline(h=qnorm(0.975),lty="dashed")
curve(qt(0.75,df=x-2), add=TRUE, col="orange")
abline(h=qnorm(0.75), lty="dashed", col="orange")
legend("topright", legend=c(expression(alpha==0.01), expression(alpha==0.05),
                            expression(alpha==0.5)),
       col=c("blue","black","orange"), lty="solid")
```

FIGURE 3: *Convergence of $k(n, \alpha)$ as $n \to \infty$, illustrated for $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.5$. (Why do I plot the 97.5^{th} percentile when I'm interested in $\alpha = 0.05$?)*

4 Statistical Significance: Uses and Abuses

4.1 p-Values

The test statistic for the Wald test,

$$T = \frac{\widehat{\beta}_1 - \beta_1^*}{\widehat{\text{se}}\left[\widehat{\beta}_1\right]}$$

has the nice, intuitive property that it ought to be close to zero when the null hypothesis $\beta_1 = \beta_1^*$ is true, and take large values (either positive or negative) when the null hypothesis is false. When a test statistic works like this, it makes sense to summarize just how bad the data looks for the null hypothesis in a **p-value**: when our observed value of the test statistic is T_{obs}, the p-value is

$$P = \mathbb{P}\left(|T| \geq |T_{obs}|\right)$$

calculating the probability under the null hypothesis. (I write a capital P here as a reminder that this is a random quantity, though it's conventional to write the phrase "p-value" with a lower-case p.) This is the probability, under the null, of getting results which are at least as extreme as what we saw. It should be easy to convince yourself that rejecting the null in a level-α test is the same as getting a p-value $< \alpha$.

It is not too hard (Exercise 4) to show that P has a uniform distribution over $[0, 1]$ under the null hypothesis.

4.2 p-Values and Confidence Sets

When our test lets us calculate a p-value, we can form a $1 - \alpha$ confidence set by taking all the β's where the p-value is $\geq \alpha$. Conversely, if we have some way of making confidence sets already, we can get a p-value for the hypothesis $\beta = \beta^*$; it's the largest α such that β^* is in the $1-\alpha$ confidence set.

4.3 Statistical Significance

If we test the hypothesis that $\beta_1 = \beta_1^*$ and reject it, we say that the difference between β_1 and β_1^* is **statistically significant**. Since, as I mentioned, many professions have an overwhelming urge to test the hypothesis $\beta_1 = 0$, it's common to hear people say that "β_1 is statistically significant" when they mean "β_1 is difference from 0 is statistically significant".

This is harmless enough, as long as we keep firmly in mind that "significant" is used here as a technical term, with a special meaning, and is *not* the same as "important", "relevant", etc. When we reject the hypothesis that $\beta_1 = 0$, what we're saying is "It's really implausibly hard to fit this data with a flat line, as opposed to one with a slope". This is informative, if we had serious reasons to think that a flat line was a live option.

It is incredibly common for researchers from other fields, and even some statisticians, to reason as follows: "I tested whether $\beta_1 = 0$ or not, and I retained the null; *therefore* β_1 is *in*significant, and I can ignore it." This is, of course, a complete fallacy.

To see why, it is enough to realize that there are (at least) two reasons why our hypothesis test might retain the null $\beta_1 = 0$:

1. β_1 is, in fact, zero,

2. $\beta_1 \neq 0$, but $\widehat{\text{se}}\left[\widehat{\beta}_1\right]$ is so large that we can't tell anything about β_1 with any confidence.

There is a very big difference between data which lets us say "we can be quite confident that the true β_1 is, if not perhaps exactly 0, then very small", and data which only lets us say "we have no earthly idea what β_1 is, and it may as well be zero for all we can tell." It is good practice to always compute a confidence interval, but it is *especially* important to do so when you retain the null, so you know whether you can say "this parameter is zero to within such-and-such a (small) precision", or whether you have to admit "I couldn't begin to tell you what this parameter is".

Substantive vs. statistical significance Even a huge β_1, which it would be crazy to ignore in any circumstance, can be statistically insignificant, so long as $\widehat{\text{se}}\left[\widehat{\beta}_1\right]$ is large enough. Conversely, any β_1 which isn't exactly zero, no matter how close it might be to 0, will become statistically significant at any threshold once $\widehat{\text{se}}\left[\widehat{\beta}_1\right]$ is small enough. Since, as $n \to \infty$,

$$\widehat{\text{se}}\left[\widehat{\beta}_1\right] \to \frac{\sigma}{s_X\sqrt{n}}$$

we can show that $\widehat{\text{se}}\left[\widehat{\beta}_1\right] \to 0$, and $\frac{\widehat{\beta}_1}{\widehat{\text{se}}\left[\widehat{\beta}_1\right]} \to \pm\infty$, unless β_1 is exactly 0 (see below).

Statistical significance is a weird mixture of how big the coefficient is, how big a sample we've got, how much noise there is around the regression line, and how spread out the data is along the x axis. This has so little to do with "significance" in ordinary language that it's pretty unfortunate we're stuck with the word; if the Ancestors had decided to say "statistically detectable" or "statistically distinguishable from 0", we might have avoided a lot of confusion.

If *you* confuse substantive and statistical significance in this class, it will go badly for you.

4.4 Appropriate Uses of p-Values and Significance Testing

I do not want this section to give the impression that p-values, hypothesis testing, and statistical significance are unimportant or necessarily misguided. They're often used badly, but that's true of every statistical tool from the sample mean on down the line. There are certainly situations where we really do want to know whether we have good evidence against some *exact* statistical hypothesis, and that's just the job these tools do. What are some of these situations?

Model checking Our statistical models often make very strong, claims about the probability distribution of the data, with little wiggle room. The simple linear regression model, for instance, claims that the regression function is *exactly* linear, and that the noise around this line has *exactly* constant variance. If we test these claims and find very small p-values, then we have evidence that there's a detectable, systematic departure from the model assumptions, and we should re-formulate the model.

Actual scientific interest Some scientific theories make very precise predictions about coefficients. According to Newton, the gravitational force between two masses is inversely proportional to the *square* of the distance between them, $\propto r^{-2}$. The prediction is exactly $\propto r^{-2}$, not $\propto r^{-1.99}$ nor $\propto r^{-2.05}$. Measuring that exponent and finding even tiny departures from 2 would be big news, if we had reason to think they were real and not just noise. One of the most successful theories

in physics, quantum electrodynamics, makes predictions about some properties of hydrogen atoms with a theoretical precision of one part in a trillion; finding even tiny discrepancies between what the theory predicts and what we estimate would force us to rethink lots of physics. Experiments to detect new particles, like the Higgs boson, essentially boil down to hypothesis testing, looking for deviations from theoretical predictions which should be exactly zero if the particle doesn't exist.

Outside of the natural sciences, however, it is harder to find examples of interesting, exact null hypothesis which are, so to speak, "live options". The best I can come up with are theories of economic growth and business cycles which predict that the share of national income going to labor (as opposed to capital) should be constant over time. Otherwise, in the social sciences, there's usually little theoretical reason to think that certain regression coefficients should be *exactly* zero, or *exactly* one, or anything else.

5 Confidence Sets and p-Values in R

When we estimate a model with `lm`, R makes it easy for us to extract the confidence intervals of the coefficients:

```
confint(object, level=0.95)
```

Here `object` is the name of the fitted model object, and `level` is the confidence level; if you want 95% confidence, you can omit that argument. For instance:

```
library(gamair); data(chicago)
death.temp.lm <- lm(death ~ tmpd, data=chicago)
confint(death.temp.lm)
```

```
##                   2.5 %     97.5 %
## (Intercept) 128.8783687 131.035734
## tmpd         -0.3096816  -0.269607
```

```
confint(death.temp.lm, level=0.90)
```

```
##                    5 %        95 %
## (Intercept) 129.0518426 130.8622598
## tmpd         -0.3064592  -0.2728294
```

If you want p-values for the coefficients, those are conveniently computed as part of the `summary` function:

```
coefficients(summary(death.temp.lm))
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 129.9570512 0.55022802 236.18763  0.00000e+00
## tmpd         -0.2896443 0.01022089 -28.33845 3.23449e-164
```

Notice how this actually gives us an array with four columns: the point estimate, the standard error, the t statistic, and finally the p-value. Each row corresponds to a different coefficient of the model. If we want, say, the p-value of the intercept, that's

```
coefficients(summary(death.temp.lm))[1,4]
```

```
## [1] 0
```

The summary function will also print out a *lot* of information about the model:

```
summary(death.temp.lm)
```

```
##
## Call:
## lm(formula = death ~ tmpd, data = chicago)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.275  -9.018  -0.754   8.187 305.952
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 129.95705    0.55023  236.19   <2e-16 ***
## tmpd         -0.28964    0.01022  -28.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.22 on 5112 degrees of freedom
## Multiple R-squared:  0.1358,Adjusted R-squared:  0.1356
## F-statistic: 803.1 on 1 and 5112 DF,  p-value: < 2.2e-16
```

As my use of `coefficients(summary(death.temp.lm))` above suggests, the `summary` function actually returns a complex object, which can be stored for later access, and printed. Controlling how it gets printed is done through the `print` function:

```
print(summary(death.temp.lm), signif.stars=FALSE, digits=3)
```

```
##
## Call:
## lm(formula = death ~ tmpd, data = chicago)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42.27  -9.02  -0.75   8.19 305.95
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 129.9571     0.5502   236.2   <2e-16
## tmpd         -0.2896     0.0102   -28.3   <2e-16
##
## Residual standard error: 14.2 on 5112 degrees of freedom
## Multiple R-squared:  0.136,Adjusted R-squared:  0.136
## F-statistic:  803 on 1 and 5112 DF,  p-value: <2e-16
```

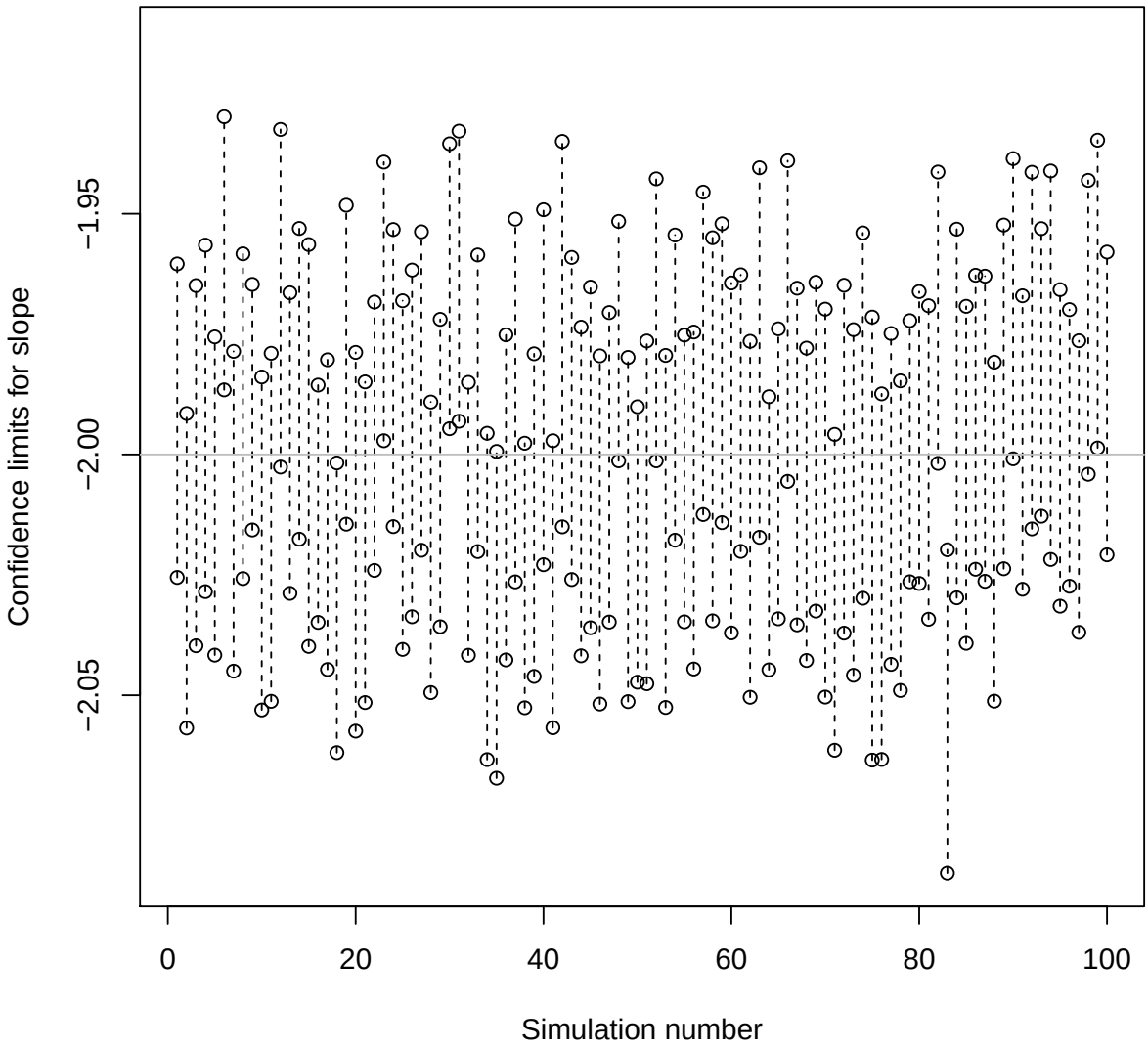
5.1 Coverage of the Confidence Intervals: A Demo

Here is a little computational demonstration of how the confidence interval for a parameter is a random parameter, and how it covers the true parameter value with the probability we want. I'll repeat many simulations of the model from Figure 2, calculate the confidence interval on each simulation, and plot those. I'll also keep track of how often, in the first m simulations, the confidence interval covers the truth; this should converge to $1 - \alpha$ as m grows.

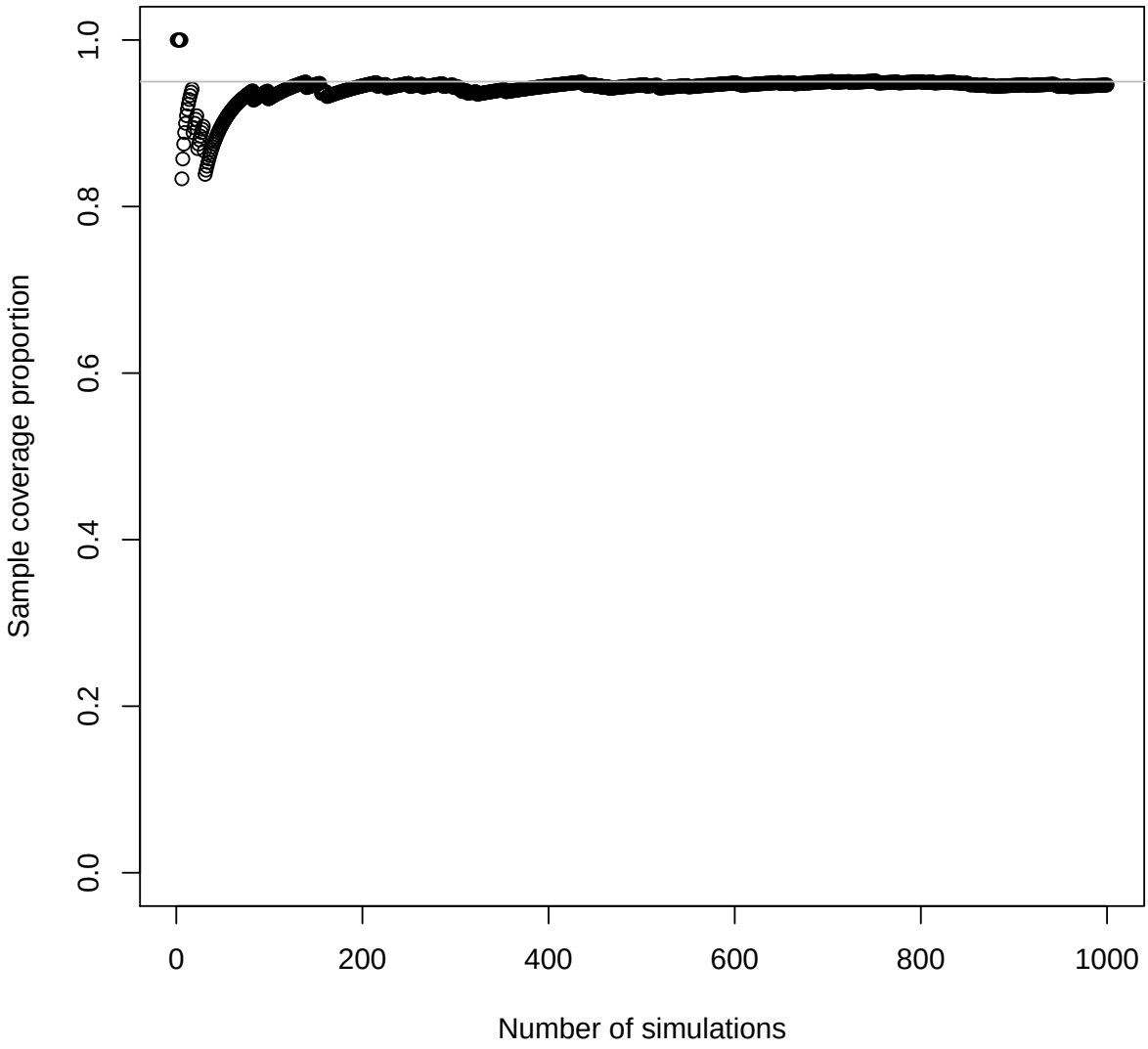
Exercises

To think through or to practice on, not to hand in.

1. *Confidence interval for σ^2:* Start with the observation that $n\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$.
   (a) Find a formula for the $1-\alpha$ sampling interval for $\widehat{\sigma}^2$, in terms of the CDF of the χ^2_{n-2} distribution, α, n and σ^2. (Some of these might not appear in your answer.) Is the width of your sampling interval the same for all σ^2, the way the width of the sampling interval for $\widehat{\beta}_1$ doesn't change with β_1?
   (b) Fix $\alpha = 0.05$, $n = 40$, and plot the sampling intervals against σ^2.
   (c) Find a formula for the $1-\alpha$ confidence interval for σ^2, in terms of $\widehat{\sigma}^2$, the CDF of the χ^2_{n-2} distribution, α and n.
2. Suppose we start a way of testing the hypothesis $\beta = \beta^*$ which can be applied to any β^*, and which has size (false alarm / type I error) probability α for β^*. Show that the set of β retained by their tests is a confidence set, with confidence level $1-\alpha$. What happens if the size is $\leq \alpha$ for all β^* (rather than exactly α)?
3. Suppose we start from a way of creating confidence sets which we know has confidence level $1-\alpha$. We test the hypothesis $\beta = \beta^*$ by rejecting when β^* is outside the confidence set, and retaining when β^* is inside the confidence set. Show that the size of this test is α. What happens if the initial confidence level is $\geq 1-\alpha$, rather exactly $1-\alpha$?
4. Prove that the p-value P is uniformly distributed under the null hypothesis. You may, throughout, assume that the test statistic T has a continuous distribution.
   (a) Show that if $Q \sim \text{Unif}(0,1)$, then $P = 1 - Q$ has the same distribution.
   (b) Let X be a continuous random variable with CDF F. Show that $F(X) \sim \text{Unif}(0,1)$. *Hint:* the CDF of the uniform distribution $F_{\text{Unif}(0,1)}(x) = x$.
   (c) Show that P, as defined, is $1 - F_{|T|}(|T_{obs}|)$.
   (d) Using the previous parts, show that $P \sim \text{Unif}(0,1)$.
5. Use Eq. **??** to show Eq. **??**, following the derivation of Eq. **??**.




```
# Run 1000 simulations and get the confidence interval from each
CIs <- replicate(1000, confint(lm(y~x,data=sim.gnslrm(x=x,5,-2,0.1,FALSE)))[2,])
# Plot the first 100 confidence intervals; start with the lower limits
plot(1:100, CIs[1,1:100], ylim=c(min(CIs),max(CIs)),
     xlab="Simulation number", ylab="Confidence limits for slope")
# Now the lower limits
points(1:100, CIs[2,1:100])
# Draw line segments connecting them
segments(x0=1:100, x1=1:100, y0=CIs[1,1:100], y1=CIs[2,1:100], lty="dashed")
# Horizontal line at the true coefficient value
abline(h=-2, col="grey")
```




```
# For each simulation, check whether the interval covered the truth
covered <- (CIs[1,] <= -2) & (CIs[2,] >= -2)
# Calculate the cumulative proportion of simulations where the interval
# contained the truth, plot vs. number of simulations.
plot(1:length(covered), cumsum(covered)/(1:length(covered)),
     xlab="Number of simulations",
     ylab="Sample coverage proportion", ylim=c(0,1))
abline(h=0.95, col="grey")
```